DOCUMENT RESUME

ED 431 017 TM 029 849

AUTHOR Wolfe, Edward W.; Gregoire, Michele
TITLE Using the Rasch Model To Assess the Implementation of Exemplary Middle School Practices: A Pilot Study of Florida's Middle Schools.
PUB DATE 1999-04-00
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Assistant Principals; *Block Scheduling; *Educational Practices; *Item Response Theory; *Middle Schools; Pilot Projects; Program Implementation; Telephone Surveys
IDENTIFIERS Florida; *Rasch Model

ABSTRACT

Researchers designed and piloted a questionnaire that measures the level of implementation of exemplary middle school practices (based on the work of P. George and W. Alexander, 1993) using Rasch measurement theory. Assistant principals (n=26) participated in telephone interviews by responding to a 27-item questionnaire that contains items reflecting school scheduling practices, team teaching, teacher planning, school philosophy, tracking, and other exemplary middle school practices. Findings show that schools with block scheduling exhibited more exemplary middle school practices than did those with traditional class scheduling. In addition, 58% of the sample exhibited exemplary middle school practices based on the most distinguishing questionnaire items. An appendix contains the study questionnaire. (Contains 3 tables, 2 figures, and 24 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 431 017

Running Head: USING THE RASCH MODEL TO ASSESS THE IMPLEMENTATION

Using the Rasch Model to Assess the Implementation of Exemplary Middle School Practices: A Pilot Study of Florida's Middle Schools

Michele Gregoire

Edward W. Wolfe

University of Florida

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Michele Gregoire

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

TM029849

BEST COPY AVAILABLE

Abstract

We designed and piloted a questionnaire that measures the level of implementation of exemplary middle school practices (George & Alexander, 1993) using Rasch measurement theory. Assistant principals (N = 26) participated in telephone interviews by responding to a 27-item questionnaire that contains items reflecting school scheduling practices, team teaching, teacher planning, school philosophy, tracking, and other exemplary middle school practices. Our results show that schools with block scheduling exhibited more exemplary middle school practices than did those with traditional class scheduling, $t(25) = 3.65$, $p < .05$, $d = .57$. In addition, 58% of the sample exhibited exemplary middle school practices based on the most distinguishing questionnaire items.

Adolescence, while perhaps not the period of storm and stress once suspected, still is a vulnerable time for middle school students. School environment, particularly the transition to junior high school, has been linked to changes in general motivation (Anderman & Maehr, 1994), intrinsic motivation and perceived competence (Harter, Whitesell, & Kowalski, 1992), and in self-esteem (Eccles, 1993). Recognizing this, policymakers in the 1970s focused their efforts on transforming junior high schools into middle schools whose purpose was to be:

> geared specifically to the social, psychological, moral, and intellectual needs of early adolescents. The school's organization, curriculum, and instruction were to help boys and girls make a smooth transition from elementary to high school while building their self-esteem and nourishing their unexplored talents (Cuban, 1992, p. 243).

This kind of stage-environment fit has been shown to be related to positive outcomes for adolescents (Eccles, 1993), but, as Cuban documented, many of these transformed schools do not live up to their intended purpose of being places that meet adolescent students' psychological needs. Consequently, it is important to be able to tell whether or not schools that call themselves middle schools are indeed embracing a new philosophy of schooling or whether they are mini-high schools or junior high schools in disguise.

Purpose of Study

Although general consensus exists about the practices and policies that constitute a good middle school (Fry & Jobe, 1996; George & Alexander, 1993; Lee & Milburn, 1994; McEwin, Dickinson, & Jenkins, 1996), many names for this construct populate the literature, from the more general concept of middle school philosophy (Eccles, Lord, & Buchanan, 1996), to responsive practices (Mac Iver & Epstein, 1991), to that of exemplary schools (George & Alexander, 1993). For the purposes of this paper, this construct will be referred to as exemplary

middle school practices as advanced by Bill Alexander (George & Alexander, 1993) who is generally considered one of the founding fathers of the middle school movement (McEwin et al., 1996). This construct refers to a set of general principles that are based on research about adolescents' developmental needs and schooling practices that support those needs as distinct from the practices of elementary and high schools. According to George and Alexander (1993), common practices exhibited by exemplary middle schools include interdisciplinary teams, block scheduling, teacher advisors, exploratory programs, shared governance, a variety of instructional practices, a transition program between other schools, and teachers trained in issues pertaining to adolescent development. These authors provide both summative and formative evaluations for middle schools to chart their progress towards implementing these goals. In addition, organizations such as the National Middle School Association (National Middle School Association, 1982), exist to promote this concept and mission of good middle schools. However, for the purposes of state and national policymaking as well as research on adolescents and schooling, it is helpful to be able to assess the degree to which a school has implemented such exemplary middle school practices with a less time-consuming measure than personal observation of the school site. It is for this purpose that surveys measuring the implementation of exemplary middle school practices have been created.

A review of the literature revealed that there have been generally two types of survey designs employed to study this construct. One is the small-scale administration, by mail, of a more general questionnaire designed to assess the degree to which the construct has been implemented within a local region of schools (Fry & Jobe, 1996; Lee & Milburn, 1994). Another is a more extensive questionnaire, usually conducted by mail with telephone follow-ups, in which larger patterns of implementation are assessed (Mac Iver & Epstein, 1991; McEwin et al.,

1996). Of these, the most influential is the research conducted by Mac Iver and Epstein in that they went beyond reporting the proportion of respondents who answered "yes" to various practices and policies by using inferential statistical techniques to determine what characteristics of middle schools are related to school and student success.

Measuring the implementation of middle school philosophy via a survey is problematic however. Each of the aforementioned studies presented school principals' self-reported responses. This could lead to biased results because the leader of a school may be more inclined than other school personnel to report that school in a favorable light. Furthermore, none of the above papers contained reliability and validity analyses for the questionnaires used. It is unclear whether all the items of each questionnaire are representative of the construct being measured. In addition, some of the question themselves were either too vague and subjective or too focused on individual items that may or may not coalesce to be representative of a single construct. An example of the former from Fry and Jobe's (1996) survey is "The middle school should be built on continuous progress taking into account learning style" (p. 34). Examples of the latter are the long checklist of items in the National Middle School Association questionnaire (McEwin et al., 1996) or the focus on only four practices (teaming, remedial programs, teacher advisors, and transition programs) in the study by Mac Iver and Epstein (1991).

To build upon and advance the work of these survey researchers, the first author developed and piloted an instrument to measure the level of implementation of exemplary middle school practices as conceptualized by Alexander and George (1993) as the basis for a larger study to assess the degree of implementation of these practices in Florida's middle schools. This paper focuses on the results of the pilot study with emphasis on reliability and validity analyses.

Questionnaire Design and Development

Using the criteria set forth by George and Alexander (1993), the first author created a 27-item questionnaire, the Middle School Practices Questionnaire, to assess the degree of implementation of exemplary school practices in Florida's middle schools, with response options ranging from 0 to 100 % implementation for items such as "Does your building have facilities that have been designed with middle school program needs in mind?" The questions were designed according to the criteria set forth by Converse and Presser (1986) and reviewed for face validity by 4 fellow graduate students in education taking a course in survey methodology. The second draft of this instrument was pretested with both the Principal (P), the usual source of information for surveys such as this one, and the Assistant Principal (AP) of a single school because it was thought that the Assistant Principal would be less invested in presenting an overly positive view of the school. In addition, the school was one with which the first author was familiar, so it was possible to check the accuracy of responses with experiential observation of the school. The results of this pretest revealed that the Principal did give a more positive slant to the school than did the Assistant Principal (16 responses of "fully implemented" versus 13 given by the AP). Moreover, the pretest revealed that asking the respondents to judge the implementation of practices on a scale of 0 to 100 % made the evaluatory aspect of the questionnaire more salient than desirable, especially in a school-based context where grades are how performance is judged. The final problem revealed by the pretest was that items that were more specific and factual (e.g., the presence of intramural sports) were more likely to show less than 100 % implementation than those questions that were more vague or subjective in nature (e.g., having a school philosophy based on adolescent needs).

After analyzing these results, a second draft of the questionnaire was created that differed

from the previous one in the following ways: (a) the questions became more factual and specific (e.g., a teacher advisor program exists to match a teacher to every student) and (b) the response options were changed from degree of implementation to either dichotomous choices (e.g. either agree or disagree) or partial credit responses (such as weekly, daily, or monthly planning time for teachers). Wanting to avoid sampling from our population to preserve the availability of all schools for the pilot, the second draft was pretested via the telephone with a teacher who taught at an exemplary middle school in a different state. This second stage of the pretest yielded minor changes in questionnaire wording, such as adding the phrase "at your school" to some of the questions that needed this modifier, as well as changing the agree/disagree responses to yes/no to emphasize that factual responses were being sought.

At this point, Paul George, one of the principal voices in the effort to define and evaluate exemplary middle schools (George & Alexander, 1993), reviewed the questionnaire for content validity. He said that the questionnaire reflected the underlying concept of exemplary middle school practices and made suggestions for two changes in the item wordings (P. George, personal communication, February 9, 1998). The first was that an item concerning student grouping should refer to a specific subject such as mathematics, and the second was that a question on positive climate should be made more specific because it was too subjective. In addition, he supported the decision to survey assistant principals because, to his knowledge, it had not been done before, and he seconded the hypothesis that assistant principals typically have less invested in presenting a certain view of their schools than do the principals.

Method

The final draft of the questionnaire that was used in the pilot study is presented in the Appendix. The question on grouping (#16) was changed, as per Paul George's recommendation,

to focus on math classes, and the positive climate question was eliminated because after further review, it was determined that it was a poor choice for obtaining accurate self-report data. Onsite evaluation of the school, or interviews with a sample of parents, students, and teachers would be much better indicators of whether or not school climate was positive. The dichotomous coding for each question reflects the final coding schema chosen. Originally, the data were scaled to a partial credit Rasch model (Wright & Masters, 1982) using FACETS (Linacre, 1993). However, many of the scale requirements (Linacre, 1997) were not met, so the items were recoded to dichotomous responses. Data were recoded based on the guidelines suggested by Wright and Linacre (1992). For instance, for the question about teacher planning time, 77 % said that teachers had planning time daily, 11 % said that planning time was not offered, and only one assistant principal (3 %) chose the partial credit response where planning time was offered weekly (8% did not answer this question). After reviewing the literature on middle schools, it made sense to consider those schools that offered daily planning time to teachers as more truly representative of exemplary middle schools than those that only offered it weekly or not at all. Furthermore, it seems that for schools that offer planning time to teachers at all (#18), they are more likely to report that the planning time occurs daily rather than less frequently. Responses that reflected implementation of exemplary middle school policies and practices received a one and those that did not received a zero during the coding process.

The first analysis that was run after the partial credit model was abandoned was a FACETS analysis (Linacre, 1993) of persons and items using the dichotomous model that is written as

$$\phi_{ni} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)} \qquad (1)$$

where ϕ_{ni1} is the probability that an AP selected 1 rather than 0 on item i, β_n the assistant principal's ability to interpret the school's extent of implementation of exemplary middle school practices, and δ_{i1} the difficulty of scoring 1 on that item (Wright & Masters, 1982). This analysis yields a model that provides a log-linear measure for dichotomous data so that statistical methods of data analysis can be used with categorical data. In addition, it provides useful equal-interval measures of reliability and validity to ascertain if the collected data fit the specified model. For the purposes of this pilot study, Rasch measurement yields a significant advantage over other procedures for fine-grained investigation of individual persons and items in survey research, and it is the predominant means of analysis used in this report.

Finally, the survey interviews were conducted by telephone because of concern about low response rates associated with mailed questionnaires (Czaja & Blair, 1996). Because this was a pilot study with a small, purposive sample of schools ($\underline{N}$ = 34), getting enough assistant principals, who are usually quite busy with discipline and other pressing school issues, to complete the questionnaire was a top priority. To help collect the data, three additional interviewers were trained to administer the questionnaire; however, the first author was responsible for collecting data for the majority of the schools (74 %). An advance letter was sent to assistant principals 2 weeks before the data collection process in order to inform them of this study. Assistant principals were told that their middle school's policies and practices were of interest, but they were not informed of the concept being assessed. One problem that arose from this process was that the names of the schools' assistant principals, unlike the principals' names, were not identified in any documents provided by the state. Therefore, the letters were addressed to a generic "Assistant Principal." This resulted in some assistant principals not receiving the advance letter. Furthermore, because most of the schools in the sample had more than one

assistant principal, interviewers were not sure of which assistant principal they should contact. In the future, if assistant principals cannot be identified prior to the study, then the envelopes as well as the letters should clearly state to share this letter with all the assistant principals involved in curriculum and instruction, and interviewers should specify that they want to speak to the assistant principal in charge of curriculum and instruction.

Norms and Sampling

Public schools in Florida that are self-identified as middle schools were the target population of this study. Because we were interested in the degree to which public schools that call themselves middle schools are implementing known exemplary middle school practices, public junior high schools were excluded from the study. For the aims of this pilot study, because the sample size was small, a purposive sample of schools was chosen to be representative of the larger population (Kalton, 1983). The state was divided into 12 regions based on geographic and population density representativeness. From each area, two to four schools were randomly chosen as part of the sample. The sampling frame employed was a list of all the middle and junior high schools in the state provided by the Florida Department of Education upon request. However, because Lee and Milburn's (1994) study of the implementation of middle school concepts suggested that rural schools face different problems with such implementation than do other schools, a majority of schools were selected from the 10 counties with the greatest population density in Florida (Florida Department of Commerce, 1995). In addition, seven schools that George and Alexander (1993) identified as exemplary were chosen for validity comparisons and will be discussed in that section of the paper.

Out of 34 schools selected for the sample, 26 assistant principals responded to the survey (response rate = 77 %). Assistant principals chose not to respond for two reasons: (a) they were

too busy (n = 6) or (b) there was no assistant principal at the school (n = 2). Three of the assistant principals that were too busy were in the same region of the State and were being interviewed by the same person. Otherwise, there were no other similarities in the schools that did not respond. Of the eight nonresponding schools, two were identified as exemplary, five were from the southwest region of the State, two from the north, and one from the central part of the state. We decided not to call back the respondents who said they were too busy because they had been called daily for 5 to 10 days, and we did not want to disturb them further at that point. Ultimately, even for a pilot study, a larger sample should have been selected given that the response rate was only 77 % even with diligent and repeated calling and was less than we had anticipated.

Table 1 contains descriptive statistics for the sample. Two separate one-way ANOVAs were run to compare the average survey scores for schools by location and then by region. The test for differences among the different locations was nonsignificant, F (3, 22) = .75, p = .5337, so there was no evidence that location (rural, suburban, urban, inner city) mattered. However, this analysis was based on a small, purposive sample, not a probability sample, so the results could be found to be significant in a larger, random sample of schools. Furthermore, because sampling rural schools was purposely avoided, the results should not be extended to those schools, especially in light of the evidence that there are differences in rural schools' implementation of these practices (Lee & Milburn, 1994). The test for differences among the geographic regions was also nonsignificant, F (3, 22) = 1.2, p = .3319, but may have proven significant with a larger, random sample of schools. The average minority composition of the sample of schools, 49 %, was slightly higher than that of Florida's population of students from pre-Kindergarten to 12th grade, or 43 % (Florida Department of Education, 1996).

Even though the sample was reasonably representative of the target population and there was not evidence that the average scores among regions and locations of schools differed significantly from each other, the small, nonprobability nature of the sample does not allow us to make good inferences as to what results might be found in the target population. Nevertheless, because the primary purpose of this pilot study was to examine issues of reliability and validity of the instrument itself, these results are still helpful in planning a future study. Specifically, a future study of the target population should ensure that schools are oversampled because not all schools had assistant principals and of those that did, a substantial number were too busy to participate in this study.

Reliability Analyses

Internal Consistency and Separation

As mentioned earlier, previous surveys purporting to measure a school's implementation of exemplary middle school practices have not reported how well their surveys consistently measure this underlying concept. This paper addresses that omission by analyzing person measures and item measures for internal consistency, or the sense that items are homogenous and measure the same underlying concept (Litwin, 1995). Internal consistency was measured in several ways, each with slightly different interpretations. First, Cronbach's alpha was calculated, yielding a reliability coefficient of .59 for persons. Because this is low reliability for the instrument, the point biserial correlations were examined for each item to ensure that the items were being interpreted and scored correctly. In other words, items should have a positive correlation with the underlying concept, so that answers coded as correct were associated with more of the exemplary middle school concept rather than less of it. The FACETS analysis revealed that although most of the questionnaire items had a positive correlation with total

exemplary middle school score, items 2 and 14, PE and sports respectively, had negative correlations with the concept (-.28 and -.25 respectively). According to the literature however, offering regular PE classes and intramural sports are both representative of good middle schools whereas not offering frequent PE classes or emphasizing interscholastic sports over intramural sports is not reflective of this concept (George & Alexander, 1993). Thus, we decided to eliminate these two items due to their poor fit with our model and to reanalyze the results. The results of this analysis are presented in Table 2 and reveal improved point biserial correlations for items: None were negatively associated with the underlying concept. In order to recheck the consistency of the person measures, Cronbach's alpha was recalculated on the survey data with the two items dropped (this will be the data referred to throughout the rest of this paper); reliability increased from .59 to .70, a notable increase and adequate for using this instrument for research purposes.

With regard to Rasch depictions of reliability, the separation index for principals and assistant principals was 1.51, indicating that there are 2.35 statistically distinct strata of principals (Fisher, 1992). This is supported by the reliability of separation, which equals .70, and the chi-square test of the null hypothesis of a fixed person measure: $\chi^2(25)=72.9$, $p=.00$. For the items, the separation index was 3.04, indicating that there are 4.39 statistically distinct strata. The reliability of separation for the items was .90, and the chi-square test of the fixed item difficulty null hypothesis was statistically significant: $\chi^2(15)=120.6$, $p=.00$).

Regarding the reliability of using different interviewers to collect data, because only 19% of the scores were obtained from raters other than the principal researcher, this was not as much of an issue as it might have been with more interviewers. In addition, the design did not afford an opportunity to investigate interviewer effects. That is, there was no overlap between raters with

respect to the principals or assistant principals who were interviewed. Therefore, if this pilot study is extended to a larger sample with more interviewers, then interrater reliability is something that should be taken into consideration through the use of Cohen's Kappa (Huck & Cormier, 1996) or a multi-faceted Rasch analysis (Linacre, 1993). For the purposes of this study, steps were taken to ensure that procedures were standardized for all interviewers. Each interviewer was given a fixed telephone script to follow during the telephone survey. Pretesting ensured that most of the response options were unequivocal. Even with these precautions, a few questions were still left open to interpretation by the interviewer. When this happened, interviewers wrote the unclear responses in the margins of the questionnaire and left the coding for the principal investigator to clarify. Question 2, which asked about the amount of physical education classes offered to students, is an example of this problem. The correct response option of "3 or more days a week" did not fit with those schools that had block scheduling because PE was often offered on alternating days of the week. Thus, students might take PE 3 days one week, while during the next they would only have it for 2 days. This variation may have contributed to the negative point biserial correlation mentioned previously that caused us to remove the item on PE (#2) from the data analysis.

Model Fit

In order to evaluate the fit of the data to the Rasch model, we examined the mean square fit indices for principal measures. Table 3 contains the transformed measures for each assistant principal's score, arranged in descending order. The table contains the assistant principal's measure, the standard error of that measure, a 95% confidence interval around the measure, and the standardized mean square weighted (infit) and unweighted (outfit) statistics. An examination of these standardized mean square fit indices reveals no misfit among the assistant principal

ratings (i.e., an absolute value greater than 2). The standard deviation of the infit and outfit statistics were 1.2 and .7 respectively, indicating that, as a group, the ratings of the assistant principals conform to the expectations of the Rasch model. An examination of items likewise yielded good measures of fit. The standard deviation of the standardized mean square infit and outfit statistics was .8 (for both), again, indicating good fit to the Rasch model.

Validity Analyses

Reliability is a necessary but not sufficient condition to establish the validity of a measurement instrument (Huck & Cormier, 1996). The previous section suggested that the Middle School Practices Questionnaire is measuring something consistently. The purpose of this next section is to analyze whether the something that is being measured accurately reflects the concept of interest, implementation of exemplary middle school policies and practices. Evidence for the content and construct validity of this questionnaire will be considered. Related validity issues will also be taken into account. Consequential validity will be addressed in the discussion section that follows this one.

Content Validity

Evidence in support of content validity was presented in the previous section on questionnaire design. To briefly reiterate, four experts on exemplary middle schools, a teacher, a principal, an assistant principal, and a university-based researcher reviewed this survey to ensure that the items represented the breadth and depth of the concept of interest. In addition, Paul George, the university expert on middle schools, suggested that the items on class length (item 13) and block scheduling (item 19) would be some of the best indicators of implementation of exemplary middle school practices and among the more difficult items to endorse. Figure 2 shows that these two items were harder to endorse than 67% of the other items on the survey. In

addition, Dr. George noted that the question on math courses would be the most difficult item for assistant principles to endorse due to the almost universal reliance on tracking in upper middle school math classes (P. George, personal communication, February 9, 1998 and March 25, 1999). This judgment is substantiated by the measures reported in Table 2, where item 16 on math courses (along with item 11 on policy) was one of the two most difficult items for assistant principles to endorse in this study, lending further support to the content validity of the Middle School Practices Questionnaire.

Evidence in support of substantive validity can be found by examining the point biserial correlations between item measures and scores. Positive correlations reflect that more of an item corresponds to a higher score (Linacre, 1993), which should be the case if items are theoretically sound and coded correctly. Table 2 presents the point biserial correlations for all of the included items. Further indication of substantive validity is provided in Figure 2. Items that were easier to endorse are those that are easier either to implement or are more subjective and thus easier to get correct. For instance, Question 20 asked assistant principals if a professional counselor was available to all students on a regular basis. Because school counselors are mandated by law, this question was easy to endorse. However, this question does not reveal whether or not a counselor is readily available and an integral part of the school environment, which more accurately distinguishes exemplary middle schools (George & Alexander, 1993), but which is difficult to reliably ascertain by self-report. On the other hand, as previously noted, the items on math courses and policy were quite difficult to endorse. Feedback from the principal interviewed in the pretest suggested that item 11, concerning whether school policy is set by the state and school district rather than the individual school community, is difficult to endorse because it is something over which individual schools have no control. Much public school policy is simply

handed down from above. Therefore, although the notion of local school governance is reflective of an ideal middle school, this item will very difficult to endorse, even for schools that engage in many of the other practices representative of exemplary middle schools.

Construct Validity

Even if two instruments measure the same construct, this is still not evidence that this construct is independent of other constructs. Measures of construct validity are used to make this distinction by showing that a particular construct has a strong relationship with certain theoretically relevant variables (convergent validity) while at the same time possessing a weak relationship between conceptually unrelated variables (divergent validity) (Huck & Cormier, 1996). In this study we attempted to establish construct validity in several ways. First, several schools in the sample were selected on the basis of their having been previously identified as exemplary middle schools about 5 years previously by George and Alexander (1993). It was predicted that these identified exemplary middle schools would show a strong relationship to high scores on this questionnaire and that average total scores of identified exemplary middle schools would be larger than average total scores of non-identified middle schools. The correlation for the first prediction was $r = .11$, which is a slight positive relationship. However, a t test of this correlation was non-significant, $t(24) = .54$, $p = .5941$. In addition, we failed to reject the null hypothesis that there was no difference between the means, $t(25) = .6$, $p = .5539$. Because the size of the identified group was quite small ($n = 5$), there was not enough power to detect differences that may have existed between the two groups. Furthermore, in 5 years, some identified schools may have declined in this construct while non-identified schools may have increased their score on this construct. In the future, it would be important to retest this hypothesis because it is relevant to the consequential validity of George and Alexander's roster

of identified schools. If some schools are coasting on their reputations while other good but non-identified schools are ignored, then this reflects an unfair bias that need to be corrected.

Another way that convergent validity was measured was to compare schools on scheduling issues that the university expert claimed were most representative of exemplary middle schools. Item 13 asked assistant principals about the length of their school's class periods while item 19 sought to determine if block scheduling was used. These two items were significantly correlated as predicted, $r = .57$, $t(24) = 3.39$, $p = 0.0024$. Schools were then divided into two groups: those that scored correctly on one or both items and those that received a zero on both questions. A one-tailed t test revealed that schools who had class periods longer than an hour and/ or schools that have four periods per day (block scheduling) scored higher on the implementation of exemplary middle school practices than did the group that did not implement these class schedule practices, $t(25) = 3.65$, $p = .0012$. The standardized difference between the means, or Cohen's effect size, was .57 and represents a moderate, or visible effect (Huck & Cormier, 1996).

In order to ensure that exemplary middle school practices were not associated with variables that are conceptually unrelated to this construct, the data were analyzed for evidence of discriminant validity. Both teacher/student ratio and SES level of the student population may be related to school functioning; however, these indicators are not theoretically connected to the implementation of middle school concepts. Correlations between these variables and total scores, r's of .22 and -.17 respectively, were nonsignificant, $t(24) = 1.1$ and $-.85$, $p = .2822$ and .4037 consecutively. Although this suggests that the survey provides evidence in support of discriminant validity, two issues must be kept mind. First, as mentioned previously, because the sample size was small, power to detect differences was low. Second, and more important, the

sample chosen was purposive and not random, so the results may not hold for a larger, random sample of Florida's middle schools.

Related Validity Issues

Three other validity issues are relevant to this discussion. One is a matter of external validity and concerns the initial decision to interview assistant principals rather than principals because of our hypothesis that principals' responses would show a positive response bias. Although we intended to get measures from five principals to test this hypothesis, only two principals were available to be interviewed. As expected, both assistant principals' and principals' item responses were positively and significantly correlated, $r = .57$ and .62 respectively. Furthermore, the average score for each principal was 10 and 12 while the average score of their assistant principals were 8 and 11 respectively, which suggests that assistant principals are less positively biased than are principals. In order to test statistically whether this would be true in the population, a larger, random sample of both principals and assistant principals would be necessary.

Regarding the generalizability of these findings, even if the results of this survey were generalizable to Florida's middle schools, they should not be interpreted as being representative of other practices in other states due to the local and state control of schooling that exists in the United States. Additionally, even though there was no evidence that average schools scores differed based on the location of the schools (rural, urban, suburban, or inner city) or region of the state (Central, Northwest, Southwest, or Southeast) as reported in the section on norms and sampling, further testing is needed to appropriately support this claim.

Discussion

Overall, this analysis presented evidence that the Middle School Practices Questionnaire

reliably and validly measured the implementation of exemplary middle school characteristics in this sample. As to whether or not inferences can be made about the larger population of Florida middle schools is a question that will be better answered in a full test of the questionnaire by randomly sampling a larger group of schools. Nevertheless, this pilot study has generated useful data to improve such a future study, thus achieving its primary purpose.

The theoretical question that is at the heart of this questionnaire is to what degree Florida's middle schools are implementing exemplary middle school practices. Using the measures given by the Rasch model in equation 1, and examining distribution of items in Figure 2, a cutoff score of -.03 was assigned to distinguish between those schools that can be considered exemplary middle schools and those that have not met enough of the criteria to merit that distinction. This standard was chosen because it separated schools based on the amount of hard to endorse items to which a school subscribed while including block scheduling and longer class periods, two items that were previously shown to be an important aspect of this construct. With this criterion in place, the proportion of the sample that can be considered exemplary middle schools is 58%. A 95% confidence interval for this estimate is contained in the interval .38 to .78. This is a wide interval that will be reduced when a larger sample of schools is used. Nonetheless, this is useful place to begin comparing school measures when a full study of the population is conducted. One other theoretical aspect of this study is that it provides evidence that the construct of interest, implementation of exemplary middle school practices, is one that can be measured via a telephone survey. This provides support for further studies of this construct in this manner, and it is notable that this design is much more cost-effective than onsite evaluations of the school site.

One problematic theoretical issue is why Question 14, about whether after school sports programs were primarily intramural or inter-scholastic, was negatively correlated with the construct. The literature suggests that environments that foster social comparisons and competitiveness are antithetical to adolescent needs (Eccles, 1993, McEwin et al., 1996); consequently, schools that emphasize noncompetitive team sports should be more reflective of sensitivity to adolescent needs and hence exemplary middle school practices. In future studies, this item should be rephrased so that it is not an either/or choice; instead, it should attempt to measure how strongly intramural and extracurricular sports are emphasized at the school in comparison to interscholastic sports. Perhaps a simple count of the number of students participating in each type of activity would be a more accurate measure of this item.

There are many practical implications suggested by the results obtained in this study that may generalize to other surveys of this construct in other states. One is that factual questions, such as class length or number of math courses, are better items to use in a telephone survey like this one rather than subjective ones in which respondent bias may positively skew the results. In addition, this survey has provided some evidence in support of interviewing assistant principals rather than principals in order to obtain less biased responses; this is not the usual method employed in surveys like this one, so it is a notable consideration. Unfortunately, there was not enough evidence to test this hypothesis, so it remains a mere suggestion. A further recommendation would be to be cautious with the use of partial credit answers or with responses that ask participants to state the degree of implementation of these constructs. The latter design has been used in other surveys (Fry & Jobe, 1996), but the analyses in this study did not support either of these options.

One practical issue that the interviewers faced concerned the absence of "neither" or "some" response options. This was particularly salient for the previously discussed PE item, as well as for the items on bells (#9) and curriculum orientation (#12). The latter item seemed to reflect the tension between respondents' wish to say their schools curriculum was student-oriented and the more realistic response that it was really subject-oriented; thus providing a "neither" or "a little of both" option may provide a less frustrating choice for assistant principals while still allowing those schools that are strongly student-oriented to be recognized. The item on bells revealed that some schools have bells that begin and end the day but that do not signal the end of classes. Having a school that is not totally dominated by bells is more in line with exemplary middle school practices than schools where bells signal the end and beginning of all classes; consequently, the response options for this item should be modified to include a choice that gives credit for the situation where bells signal certain major breaks in the day, but not breaks between classes. Given a larger sample of respondents, Rasch's partial credit model (Wright & Masters, 1982) might be able to distinguish between such levels of implementation. A final suggestion to improve the questionnaire concerns item 18, planning time for teacher teams. Because assistant principals were likely to say their teachers had daily (or block) team planning time if they had teams at all, the response options should be changed to accommodate block scheduling, as with the PE item, and, more important, the item should be phrased so as to inquire into the nature of this planning time. Some schools offer teachers both individual and team planning time; these are the schools that are most representative of exemplary middle schools (George & Alexander, 1993). Others schedule individual teachers' planning time at the same time for all members of a team; this is less representative of the construct, but is still better than those schools who report that team planning time is offered daily, but, in fact, mean that a team

is encouraged to meet before and after school but is not provided with a specific time in the schedule for team teachers to meet.

In addition to piloting this instrument with a larger, random sample of middle schools to obtain more stable population estimates, future reliability and validity studies should investigate whether scores on the Middle School Practices Questionnaire are consistent over time and demonstrate concurrent validity with similar measures of exemplary middle schools. A powerful means for conducting the first study would be to ascertain whether changes in school scores are the results of changes in raters, changes in items, or changes in the school by creating a common frame of reference for interpreting measures that are taken at different times through the use of Rasch measurement techniques (see Wolfe & Chiu, 1997 for a step-by-step procedure for how to accomplish this). Secondly, evidence in supportive of concurrent validity for using this instrument as a valid indicator of the degree to which a middle school is engaged in exemplary practices could be obtained by correlating scores on this instrument with scores on another instrument that measures similar concepts, such as Mac Iver and Epstein's (1991) detailed survey of the use of responsive practices by middle schools. Another, more labor-intensive option would be to perform on-site summative evaluations of selected exemplary middle schools as explained by George and Alexander (1993). Positive, high correlations between the two measures would establish that both instruments are measuring the same concept.

A few final words about the consequential validity of this study are in order. Surveys such as this one may be used to evaluate schools for the purposes of merit pay and other benefits, or they may be used to penalize schools for failing to achieve a certain ideal. Both of these purposes are not well-served by the use of a survey such as this one. Not only is the reliability of this instrument too low for such high stakes decisions (alpha = .70), but more significantly,

summative evaluations of schools are not fair if there is not a concomitant formative evaluatory process established to help support those schools that are in need of improvement. The most reasonable use of an instrument as the present one would be to contribute to the advancement of research on characteristics of schools that are related to positive learning outcomes for students or to obtain rough estimates of the degree of implementation of this construct in a state or across the nation (with concomitant pilot testing to ensure that the items are reliable and valid for those populations as well). With respect to the first suggestion, further validity analyses of the present instrument are needed. In order to contribute to theory-building on the characteristics of middle schools, the next step would be to conduct causal comparative analyses in which schools' scores were correlated with both student achievement and mental health outcomes. It would be a significant advancement of stage-environment fit theory (Eccles, 1993) if such positive outcomes were associated with exemplary middle schools. At the same time, the practical import of such knowledge would have a consequential impact on public school policy and practices for the coming decades.

References

Anderman, E. M., & Maehr, M. L. (1994). Motivation and schooling in the middle grades. Review of Educational Research, 64, 287-309.

Converse, J. M., & Presser, S. (1986). Survey questions: Handcrafting the standardized questionnaire. (Vol. 07-063). Thousand Oaks, CA: SAGE.

Cuban, L. (1992). What happens to reforms that last? The case of the junior high school. American Educational Research Journal, 29, 227-251.

Czaja, R., & Blair, J. (1996). Designing surveys: A guide to decisions and procedures. Thousand Oaks, CA: Pine Forge Press.

Eccles, J. S. (1993). School and family effects of the ontogeny of children's interests, self-perceptions, and activity choices. In J. Jacobs (Ed.), Developmental perspectives on motivation. Nebraska Symposium on Motivation, Vol. 40 (pp. 145-208). Lincoln: University of Nebraska Press.

Eccles, J. S., Lord, S., & Buchanan, C. M. (1996). School transitions in early adolescence: What are we doing to our young people? In J. A. Graber, J. Brooks-Gunn, & A. C. Petersen (Eds.), Transitions through adolescence: Interpersonal domains and context (pp. 251-284). Mahwah, NJ: Erlbaum.

Fisher, W.P. (1992). Reliability statistics. Rasch Measurement Transactions, 9, 472.

Florida Department of Commerce. (1995). Population growth, 1990 to 1994 [On-line]. Available: http://www.state.fl.us/commerce/public-html/html/publications/ctycomp.html.

Florida Department of Education. (1996). PK-12 student membership by racial/ethnic category, fall 1996 [On-line]. Available: http://www.firn.edu/doe/bin00050/race.htm.

Fry, D. D., & Jobe, M. E. (1996). Importance and implementation of middle school concepts: A study of East Texas middle school principals. ERS Spectrum, 14(1), 31-36.

George, P. S., & Alexander, W. M. (1993). The exemplary middle school. (2nd ed.). Fort Worth, TX: Harcourt Brace College Publishers.

Harter, S., Whitesell, N. R., & Kowalski, P. (1992). Individual differences in the effects of educational transitions on young adolescents' perceptions of competence and motivational orientation. American Educational Research Journal, 29, 777-807.

Huck, S. W., & Cormier, W. H. (1996). Reading statistics and research. (2nd ed.). New York: HarperCollins.

Kalton, G. (1983). Introduction to survey sampling. (Vol. 07-035). Newbury Park, CA: SAGE.

Lee, S., & Milburn, C. (1994). Implementing middle school concepts in rural areas: Problems and solutions. Middle School Journal, 26(1), 7-11.

Linacre, J. M. (1993). Many-facet Rasch measurement. Chicago: MESA.

Linacre, J.M. (1997). Guidelines for rating scales. MESA Research Note #2. Chicago: MESA.

Litwin, M. S. (1995). How to measure survey reliability and validity. Thousand Oaks, CA: SAGE.

Mac Iver, D. J., & Epstein, J. L. (1991). Responsive practices in the middle grades: Teacher teams, advisory groups, remedial instruction, and school transition programs. American Journal of Education, 99, 587-622.

McEwin, C. K., Dickinson, T. S., & Jenkins, D. M. (1996). America's middle schools: Practices and progress. A 25 year perspective. Columbus, OH: National Middle School Association.

National Middle School Association. (1982). This we believe. Columbus, OH: National Middle School Association.

Wolfe, E. W., & Chiu, C. W. T. (1997). Measuring change over time. Manuscript in preparation.

Wright, B.D., & Linacre, J.M. (1992). Combining and splitting categories. Rasch Measurement Transactions, 11, 550-551.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press.

Appendix

Middle School Practices Questionnaire

Interviewers: Do not read the italicized phrases to the respondents.

<<I'm going to read a series of statements and questions about your school as it currently stands right now. Respond to each statement or question with the answer that **best** reflects your school from the choices that are given. Remember that your answers should reflect your school's **present** reality including those limitations that might be due to lack of resources or policies set at the state or local level. >>

1. Your middle school consists of grades:
a) 6-9 ..... *0*
b) 5-8 ..... *1*
c) 6-8 **or** *1*
d) 7-9 ..... *0*

2. *(dropped)* Students at your school attend physical education classes:
a) three or more days per week throughout the year *1*
b) less than three days per week throughout the year **or** *0*
c) P.E. classes are not offered throughout the year for all students *0*

3. Your school's philosophy and practices are most similar to:
a) local high schools *0*
b) local elementary schools **or** *0*
c) neither *1*

4. Alternative assessments, as opposed to tests, are used at least 50% of the time by most of the teachers at your school.
a) yes **or** 1
b) no *0*

5. A teacher advisor program exists to match a teacher to every student.
a) yes **or** *1*
b) no *0*

6. Decision-making at your school is strongly influenced by parents. Do you:
a) strongly agree *0*
b) agree *1*
c) disagree **or** *0*
d) strongly disagree *0*

7. Your school mission statement specifically mentions serving adolescent needs.
a) yes **or** *1*
b) no *0*

8. Teachers with specific middle school training are given greater preference during the hiring process than are teachers without this training.
a) yes **or** *1*
b) no *0*

9. Bells signal the end of class periods.
a) yes **or** *0*
b) no *1*

10. For the majority of your classrooms, who or what determines the curriculum that is taught, after state and district guidelines have been taken into consideration?
a) teachers *0*
b) textbooks *0*
c) teachers and students equally **or** 1
d) students and parents *0*

11. School policy is mostly determined by agencies, such as the school board or state department of education, rather than the individual schools in your district.
a) yes **or** *0*
b) no *1*

12. In general, most classes are:
a) subject-oriented **or** *0*
b) student-oriented *1*

13. The length of the typical class period for 8th graders lasts:
a) for an hour or less *0*
b) for more than an hour **or** *1*
c) it depends on the subject matter being taught* *0**

If c **is chosen, *ask:* Are there any academic classes in the regular schedule that last for 80 minutes or longer? *(If yes, ask)* Which ones?_____
a) yes *1*
b) no *0*

14. *(dropped)* After school sports programs are primarily:
a) intramural **or** *1*
b) inter-scholastic (competitive with other schools) *0*

15. Regarding the existence of formal programs that transition students from elementary school to middle school and from middle school to high school, your school:
a) Has structured programs for both ends of the transition.....*1*
b) Has structured programs for one end of the transition **or**.....*0*
c) Does not have formally established transition programs at this time.....*0*

16. How many types of math courses exist for students in your highest grade? (for example, Algebra, Pre-Algebra, et cetera)
a) four or more.....*0*
b) three.....*0*
c) two **or**.....*0*
d) one.....*1*

17. Teachers are grouped into teams that share the same students.
a) yes **or**.....*1*
b) no*.....*0*

18. *(*Skip if answer to 17 is no)* Teacher teams share a common planning time:
a) weekly.....*0*
b) daily.....*1*
c) monthly **or**.....*0*
d) less than once a month.....*0*

19. Classes are organized into:
a) four or less periods per day.....*1*
b) five or six periods per day **or**.....*0*
c) 7 or more periods per day.....*0*

20. A professional counselor is available to all students on a regular basis.
a) yes **or**.....*1*
b) no.....*0*

Demographic Data

<<Ok, that was great! Now, before I hang up and let you get on with your day, I just need to get some background information about your school. >>

22) Is your school is located in an urban, suburban, or rural location? _____ */If urban, ask:/* Would you describe it as an inner city?_____

23) How many full-time students are currently enrolled at your school? _____

24) How many full-time teachers, not including administrators or other non-teaching staff, are employed to work at your school? _____

25) Yes or no: Are you allied with any particular reform movement, either formally or informally?_____(*If yes*) Which one?_____

26) Approximately what percentage of students attending your school is a member of an ethnic minority group?_____ (*e.g. Black, Latino, Asian, etc.*)

27) Approximately what percentage of your students qualifies for free or reduced lunch?_____

<<Thank you so much for your time. If you have any further questions or comments, you can call (*me* Michele Gregoire) at xxx-xxx-xxxx.>>

Author Note

Michele Gregoire and Edward W. Wolfe, Department of Educational Foundations.

This study was conducted as part of a graduate class on survey research taught by the second author. The first author wishes to thank her fellow graduate students, especially Dawn Shuler, Brad Moulder, Mark Reed and Sarah Mendonça for their helpful critiques during the process of questionnaire development, as well as the assistant principals and principals who generously agreed to participate in this survey. In addition, she thanks Camarie Humphrey, middle school teacher par excellence for her feedback, Paul George for his expert critique of the questionnaire items, and Patricia Ashton for her proofreading of the manuscript.

Correspondence concerning this paper should be addressed to Michele Gregoire, University of Florida, 1403 Norman Hall, P.O. Box 117047, Gainesville, Florida 32611-7047. Electronic mail may be sent via Internet to mtierra@digital.net.

Table 1

Descriptive Statistics of the Middle Schools Sampled

By Location	N	M	S
Rural	1	11.00	
Suburban	9	9.78	3.27
Urban	6	8.67	1.97
Inner City	10	10.80	2.90
By Region			
North	5	11.40	2.97
Central	6	10.50	2.07
Southwest	6	8.33	3.01
Southeast	9	9.89	2.93

Table 2

FACETS Measures, Standard Errors, and Fit Statistics for Items

Item # and Description	Measure	Model SE	Infit	Outfit	Point Bis.
11 Policy	3.06	.77	0	0	.01
16 Math courses	3.06	.77	0	0	.34
10 Curriculum	2.56	.65	0	0	.27
12 Classes	1.09	.48	-1	-1	.60
13 Class length	1.09	.48	0	1	.11
19 Block scheduling	.87	.47	1	0	.08
4 Alternative Assessments	.45	.45	0	0	.19
6 Parents	.25	.45	0	0	.33
5 Advisor program	.05	.45	1	0	.16
9 Bells	-.15	.45	0	0	.45
3 Practices	-.78	.47	0	0	.28
8 Hiring teachers	-1.51	.52	0	0	.20
15 Transition programs	-1.51	.52	0	0	.48
18 Planning time	-2.06	.61	0	0	.45
17 Teacher teams	-2.54	.67	0	0	.47
1 Grade	-3.90	1.06	0	0	.23
7 Mission statement	(-5.21)	(1.85)	--	--	.00
20 Counseling	(-5.21)	(1.85)	--	--	.00

Note. Infit and outfit statistics are standardized z-scores. Model SE = Standard errors. Point Bis. = Point-Biserial Correlations. Items 7 & 20 were answered "correctly" by all respondents.

Table 3

FACETS Measures, 95 % Confidence Intervals (CI), and Fit Statistics for AP's

AP	Measure	Model SE	Lower CI	Upper CI	Infit	Outfit
3	2.35	0.78	0.74	3.96	0	0
12	1.79	0.72	0.31	3.27	1	0
27	1.79	0.72	0.31	3.27	1	0
14	1.30	0.68	-0.10	2.70	0	0
32	1.30	0.68	-0.10	2.70	-1	0
5	0.86	0.66	-0.50	2.22	0	0
13	0.86	0.66	-0.50	2.22	-1	0
15	0.86	0.66	-0.50	2.22	0	0
24	0.86	0.66	-0.50	2.22	0	0
29	0.86	0.66	-0.50	2.22	0	0
30	0.44	0.65	-0.90	1.78	-1	-1
4	0.02	0.64	-1.30	1.34	-1	-1
8	0.02	0.64	-1.30	1.34	0	0
10	-0.03	0.66	-1.39	1.33	-1	0
18	-0.03	0.66	-1.39	1.33	0	0
7	-0.39	0.65	-1.73	0.95	-2	-1
11	-0.39	0.65	-1.73	0.95	0	0
28	-0.39	0.65	-1.73	0.95	-1	-1
19	-0.82	0.66	-2.18	0.54	0	0

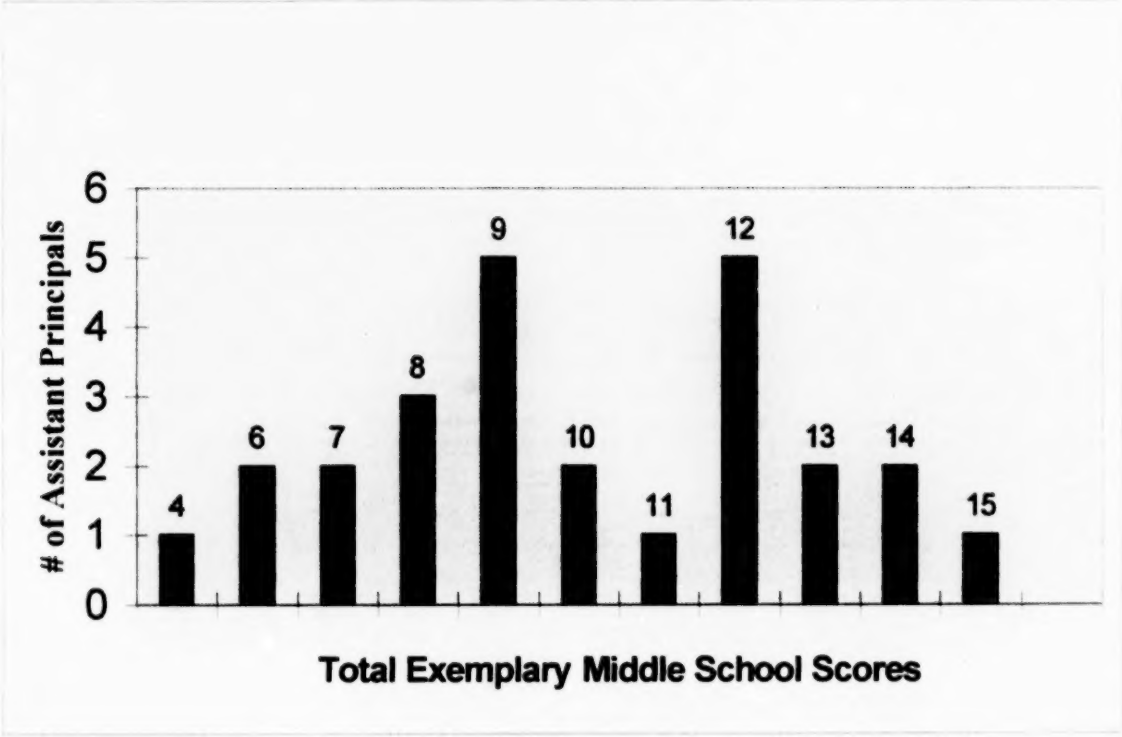
31	-0.82	0.66	-2.18	0.54	1	0
34	-0.82	0.66	-2.18	0.54	0	1
6	-1.27	0.69	-2.69	0.15	0	0
9	-1.27	0.69	-2.69	0.15	2	0
22	-1.77	0.73	-3.27	-0.27	0	0
23	-1.77	0.73	-3.27	-0.27	2	1
33	-3.03	0.90	-4.88	-1.18	0	0

Note. Infit and outfit statistics are standardized z-scores. Model SE = model standard error.

Figure Caption

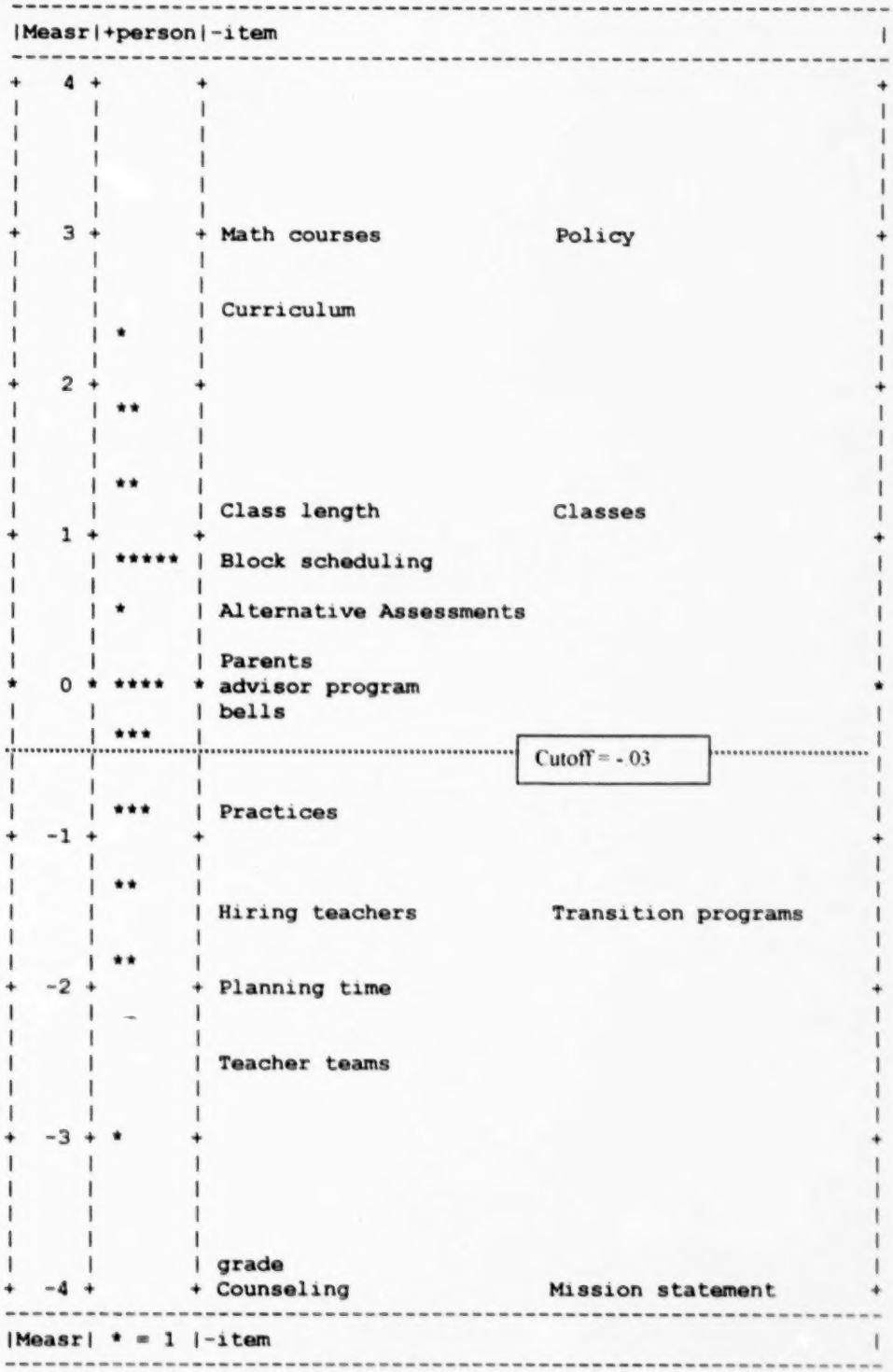
Figure 1. Frequency distribution for school scores on the Middle School Practices Survey.

Figure 2. FACETS output. Distribution of items and persons according to Rasch's dichotomous model.



BEST COPY AVAILABLE

Vertical = (1*,2A) Yardstick (columns, lines) = 0,6



BEST COPY AVAILABLE